

### Solution 1: L0 Regularization

(a) We can show

$$\arg \min_{\boldsymbol{\theta}} \mathcal{R}_{\text{reg}}(\boldsymbol{\theta}) = \arg \min_{\boldsymbol{\theta}} \sum_{i=1}^p -\hat{\theta}_i \theta_i + \frac{\theta_i^2}{2} + \lambda \mathbb{1}_{|\theta_i| \neq 0}$$

as follows:

$$\begin{aligned} \arg \min_{\boldsymbol{\theta}} \mathcal{R}_{\text{reg}}(\boldsymbol{\theta}) &= \arg \min_{\boldsymbol{\theta}} \frac{1}{2} \sum_{i=1}^n \left( y^{(i)} - \boldsymbol{\theta}^\top \mathbf{x}^{(i)} \right)^2 + \lambda \sum_{i=1}^p \mathbb{1}_{|\theta_i| \neq 0} \\ &= \arg \min_{\boldsymbol{\theta}} \frac{1}{2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2 + \lambda \sum_{i=1}^p \mathbb{1}_{|\theta_i| \neq 0} \\ &= \arg \min_{\boldsymbol{\theta}} \frac{1}{2} (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) + \lambda \sum_{i=1}^p \mathbb{1}_{|\theta_i| \neq 0} \\ &= \arg \min_{\boldsymbol{\theta}} \frac{1}{2} \mathbf{y}^\top \mathbf{y} - \mathbf{y}^\top \mathbf{X}\boldsymbol{\theta} + \frac{1}{2} \boldsymbol{\theta}^\top \mathbf{X}^\top \mathbf{X}\boldsymbol{\theta} + \lambda \sum_{i=1}^p \mathbb{1}_{|\theta_i| \neq 0} \\ &= \arg \min_{\boldsymbol{\theta}} -\mathbf{y}^\top \mathbf{X}\boldsymbol{\theta} + \frac{1}{2} \boldsymbol{\theta}^\top \mathbf{X}^\top \mathbf{X}\boldsymbol{\theta} + \lambda \sum_{i=1}^p \mathbb{1}_{|\theta_i| \neq 0} \quad (\mathbf{y}^\top \mathbf{y} \text{ does not depend on } \boldsymbol{\theta}) \\ &= \arg \min_{\boldsymbol{\theta}} -\mathbf{y}^\top \mathbf{X}\boldsymbol{\theta} + \frac{1}{2} \boldsymbol{\theta}^\top \boldsymbol{\theta} + \lambda \sum_{i=1}^p \mathbb{1}_{|\theta_i| \neq 0} \quad (\text{By assumption } \mathbf{X}^\top \mathbf{X} = \mathbf{I}) \\ &= \arg \min_{\boldsymbol{\theta}} -\hat{\boldsymbol{\theta}}^\top \boldsymbol{\theta} + \frac{1}{2} \boldsymbol{\theta}^\top \boldsymbol{\theta} + \lambda \sum_{i=1}^p \mathbb{1}_{|\theta_i| \neq 0} \\ &\quad (\text{By assumption } \mathbf{X}^\top \mathbf{X} = \mathbf{I} \text{ so that } \hat{\boldsymbol{\theta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} = \mathbf{X}^\top \mathbf{y}) \\ &= \arg \min_{\boldsymbol{\theta}} \sum_{i=1}^p -\hat{\theta}_i \theta_i + \frac{\theta_i^2}{2} + \lambda \mathbb{1}_{|\theta_i| \neq 0} \quad (\text{Writing out the inner products}) \end{aligned}$$

(b) Note that the minimization problem on the right-hand side of the previous math display can be written as  $\sum_{i=1}^p g_i(\theta_i)$ , where

$$g_i(\theta) = -\hat{\theta}_i \theta + \frac{\theta^2}{2} + \lambda \mathbb{1}_{|\theta| \neq 0}.$$

The advantage of this representation, if we are interested in finding the  $\boldsymbol{\theta}$  with entries  $\theta_1, \dots, \theta_p$  minimizing  $\mathcal{R}_{\text{reg}}(\boldsymbol{\theta})$ , is that we can minimize each  $g_i$  separately to obtain the optimal entries.

(c) Consider first the case that  $|\hat{\theta}_i| > \sqrt{2\lambda}$ . Note that

$$\begin{aligned} g_i(\hat{\theta}_i) &= -\hat{\theta}_i^2 + \frac{\hat{\theta}_i^2}{2} + \lambda & (|\hat{\theta}_i| > \sqrt{2\lambda} > 0) \\ &= -\frac{\hat{\theta}_i^2}{2} + \lambda \\ &< 0. & (|\hat{\theta}_i| > \sqrt{2\lambda} \Rightarrow -\hat{\theta}_i^2 < -2\lambda) \end{aligned}$$

Further, note that  $g_i(0) = 0$  and consequently  $g_i(\hat{\theta}_i) < g_i(0)$ . This means that 0 cannot be the minimizer of

$g_i$  in this case. Next, it holds that for any  $\theta \neq 0$  with  $\text{sgn}(\theta) = \text{sgn}(\hat{\theta}_i)$ , i.e.,  $\hat{\theta}_i\theta > 0$ , that

$$g_i(\theta) = \underbrace{-\hat{\theta}_i\theta}_{<0} + \underbrace{\frac{\theta^2}{2}}_{=\frac{(-\theta)^2}{2}} + \lambda \mathbb{1}_{|\theta| \neq 0} < \underbrace{\hat{\theta}_i\theta}_{>0} + \frac{(-\theta)^2}{2} + \lambda \mathbb{1}_{|-\theta| \neq 0} = g_i(-\theta),$$

so that for the minimizer of  $g_i$  it must hold that it has the same sign as  $\hat{\theta}_i$  in this case. Now differentiating  $g_i$  (for  $\theta > 0$ ) and setting it to zero implies

$$\begin{aligned} \frac{\partial g_i(\theta)}{\partial \theta} &= -\hat{\theta}_i + \theta \stackrel{!}{=} 0 \\ \Leftrightarrow \quad \theta &= \hat{\theta}_i. \end{aligned}$$

For sake of completeness, the second derivative is  $\frac{\partial^2 g_i(\theta)}{\partial^2 \theta} = 1 > 0$  so that we have indeed a minimum. Thus, for the minimizer  $\theta_i^*$  of  $g_i$  it must hold that  $\theta_i^* = \hat{\theta}_i$ .

(d) By taking the constraint of the case into account, we can write  $\theta_i^* = \hat{\theta}_i \mathbb{1}_{|\hat{\theta}_i| > \sqrt{2\lambda}}$ .

(e) Consider the complementary case, i.e.,  $|\hat{\theta}_i| \leq \sqrt{2\lambda}$ . As seen above it holds that  $g_i(0) = 0$ . Consider the smooth extension of  $g_i$  :

$$\tilde{g}_i(\theta) = -\hat{\theta}_i\theta + \frac{\theta^2}{2} + \lambda$$

and note that  $\tilde{g}_i$  and  $g_i$  are the same for any  $\theta \neq 0$ , while  $\tilde{g}_i$  is smooth on  $\mathbb{R}$  and  $g_i$  is discontinuous at 0. In particular, it holds that  $g_i(\theta) \leq \tilde{g}_i(\theta)$ , with equality for any  $\theta \neq 0$  and strict inequality for  $\theta = 0$ . So the minimizer of  $g_i$  is either the same as the one for  $\tilde{g}_i$  or it is zero. Similarly as above, it holds for any  $\theta \neq 0$  with  $\text{sgn}(\theta) = \text{sgn}(\hat{\theta}_i)$ , i.e.,  $\hat{\theta}_i\theta > 0$ , that

$$\tilde{g}_i(\theta) = \underbrace{-\hat{\theta}_i\theta}_{<0} + \underbrace{\frac{\theta^2}{2}}_{=\frac{(-\theta)^2}{2}} + \lambda < \underbrace{\hat{\theta}_i\theta}_{>0} + \frac{(-\theta)^2}{2} + \lambda = \tilde{g}_i(-\theta),$$

so that for the minimizer of  $\tilde{g}_i$  it must hold that it has the same sign as  $\hat{\theta}_i$  in this case. Now differentiating  $\tilde{g}_i$  and setting it to zero implies

$$\begin{aligned} \frac{\partial \tilde{g}_i(\theta)}{\partial \theta} &= -\hat{\theta}_i + \theta \stackrel{!}{=} 0 \\ \Leftrightarrow \quad \theta &= \hat{\theta}_i. \end{aligned}$$

However, if  $\hat{\theta}_i \neq 0$  then

$$\begin{aligned} g_i(\hat{\theta}_i) &= -\hat{\theta}_i^2 + \frac{\hat{\theta}_i^2}{2} + \lambda & (\text{If } \hat{\theta}_i \neq 0) \\ &= -\frac{\hat{\theta}_i^2}{2} + \lambda \\ &\geq 0, & (|\hat{\theta}_i| \leq \sqrt{2\lambda} \Rightarrow -\hat{\theta}_i^2 \geq -2\lambda) \end{aligned}$$

so that  $g_i(\hat{\theta}_i) \geq 0 = g_i(0)$ . Hence, the minimizer  $\theta_i^*$  of  $g_i$  is  $\theta_i^* = 0$ , which can be written, by taking the constraint of the case into account, as  $\theta_i^* = \hat{\theta}_i \mathbb{1}_{|\hat{\theta}_i| > \sqrt{2\lambda}}$ .

(f) In summary, we have shown that the minimizer  $\theta_i^*$  of  $g_i$  is  $\theta_i^* = \hat{\theta}_i \mathbb{1}_{|\hat{\theta}_i| > \sqrt{2\lambda}}$  for any  $i = 1, \dots, p$ . Since

$$\min_{\boldsymbol{\theta}} \mathcal{R}_{\text{reg}}(\boldsymbol{\theta}) = \min_{\boldsymbol{\theta}} \sum_{i=1}^p -\hat{\theta}_i\theta_i + \frac{\theta_i^2}{2} + \lambda \mathbb{1}_{|\theta_i| \neq 0} = \min_{\boldsymbol{\theta}} \sum_{i=1}^p g_i(\theta_i),$$

we conclude that  $\hat{\boldsymbol{\theta}}_{L0} = (\hat{\theta}_{L0,1}, \dots, \hat{\theta}_{L0,p})^\top$  given by

$$\hat{\theta}_{L0,i} = \hat{\theta}_i \mathbb{1}_{|\hat{\theta}_i| > \sqrt{2\lambda}}, \quad i = 1, \dots, p,$$

is the minimizer of the  $L_0$ -regularized empirical risk over the linear models.

## Solution 2: Regularization

(a) 

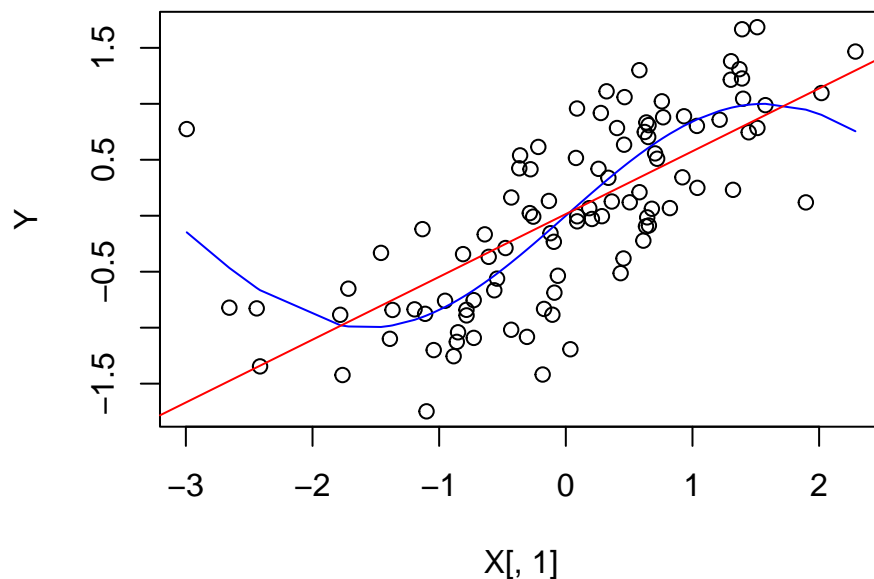
```
set.seed(42)
n = 100
p_add = 100
# create matrix of features
X = matrix(rnorm(n * (p_add + 1)), ncol = p_add + 1)

Y = sin(X[,1]) + rnorm(n, sd = 0.5)
```

(b) Demonstration of

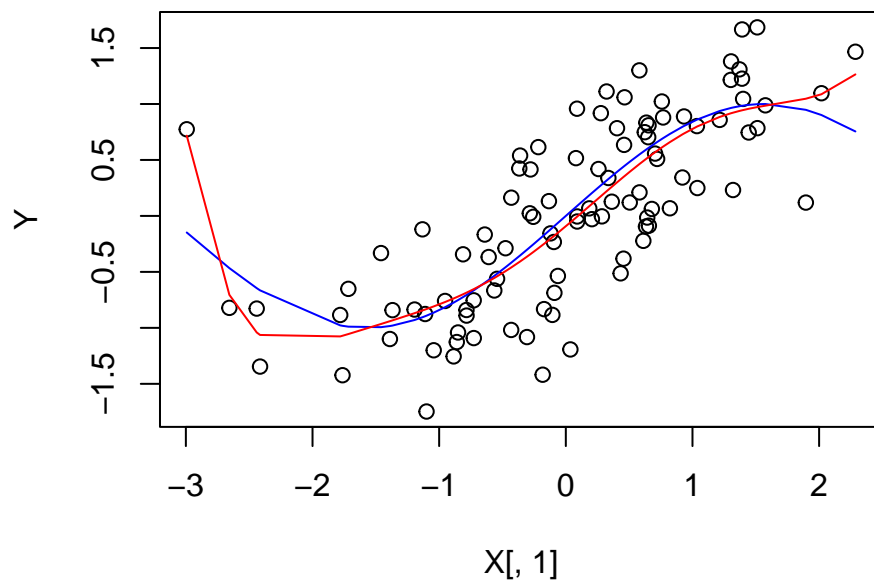
- underfitting:

```
plot(X[,1], Y)
points(sort(X[,1]), sin(sort(X[,1])), type="l", col="blue")
abline(coef(lm(Y ~ X[,1])), col="red")
```



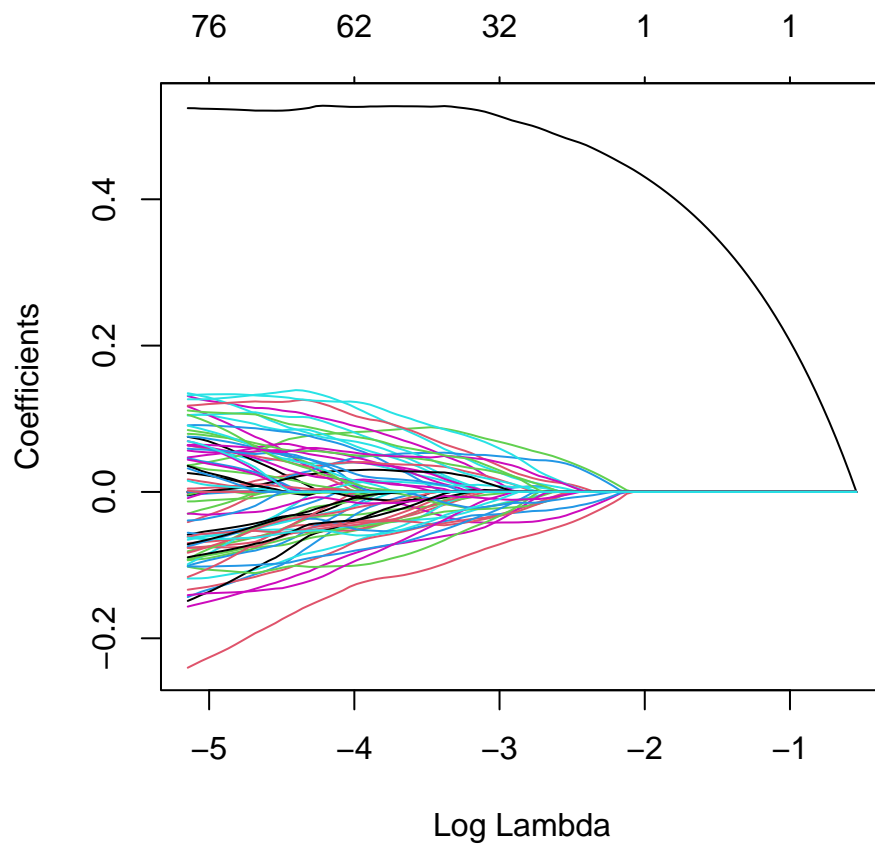
- overfitting:

```
plot(X[,1], Y)
sX1 <- sort(X[,1])
points(sX1, sin(sX1), type="l", col="blue")
points(sX1, fitted(lm(Y ~ X[,1] + I(X[,1]^2) + I(X[,1]^3) +
                     I(X[,1]^4) + I(X[,1]^5) + I(X[,1]^6) +
                     I(X[,1]^7)))[order(X[,1])],
       type="l", col="red")
```



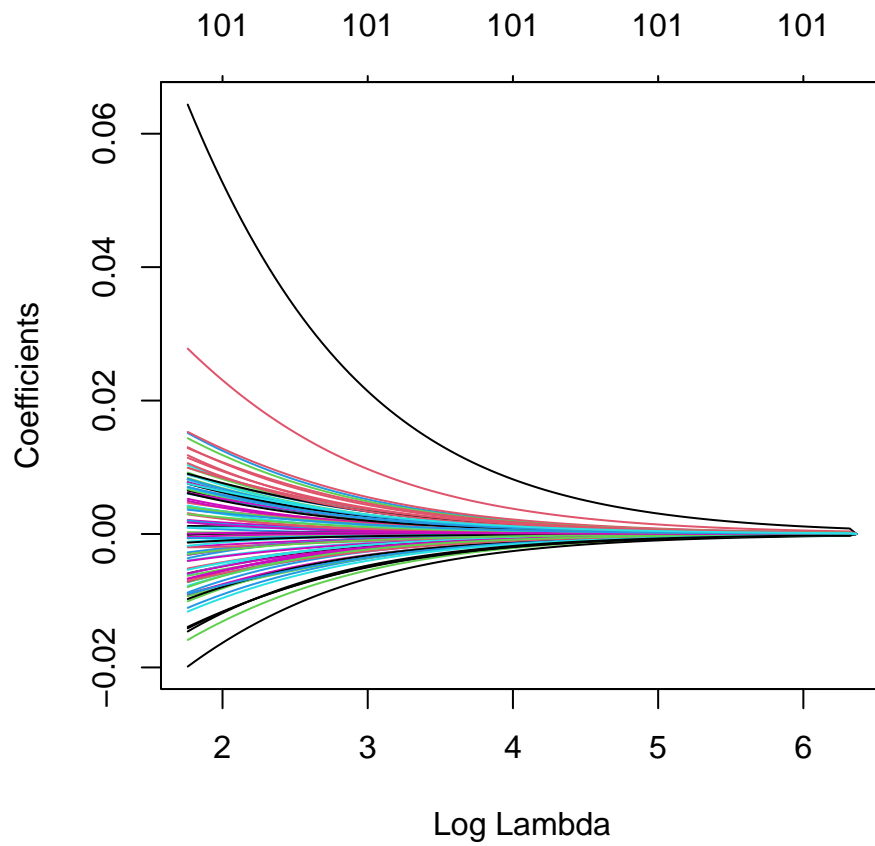
- $L1$  penalty:

```
library(glmnet)
plot(glmnet(X, Y), xvar = "lambda")
```



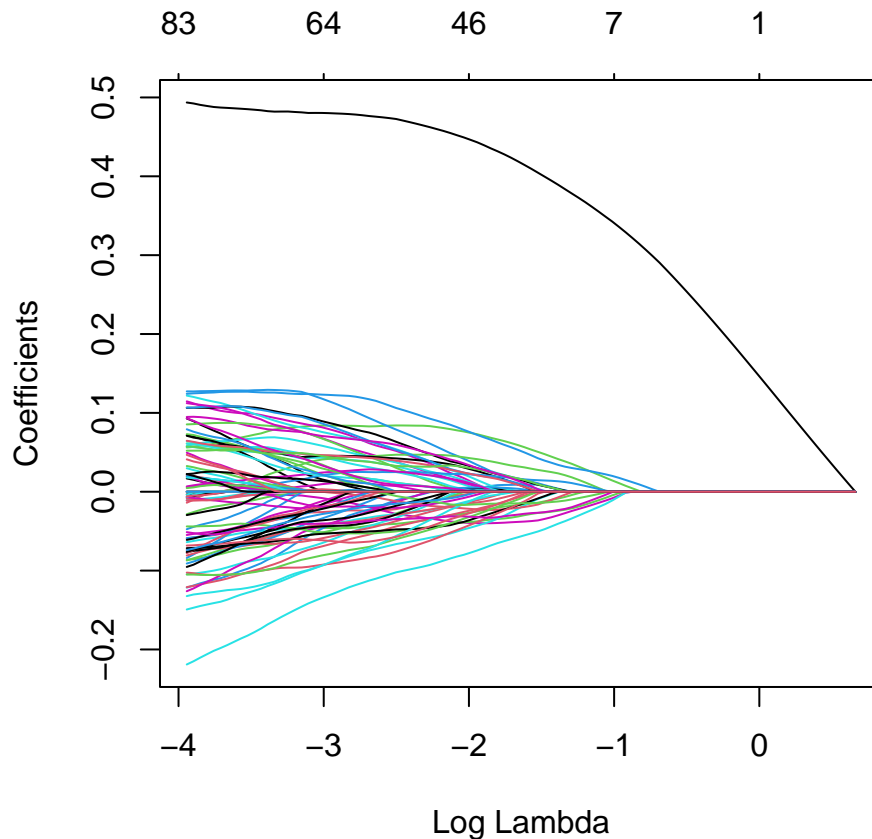
- $L_2$  penalty

```
plot(glmnet(X, Y, alpha = 0), xvar = "lambda")
```



- elastic net regularization:

```
plot(glmnet(X, Y, alpha = 0.3), xvar = "lambda")
```



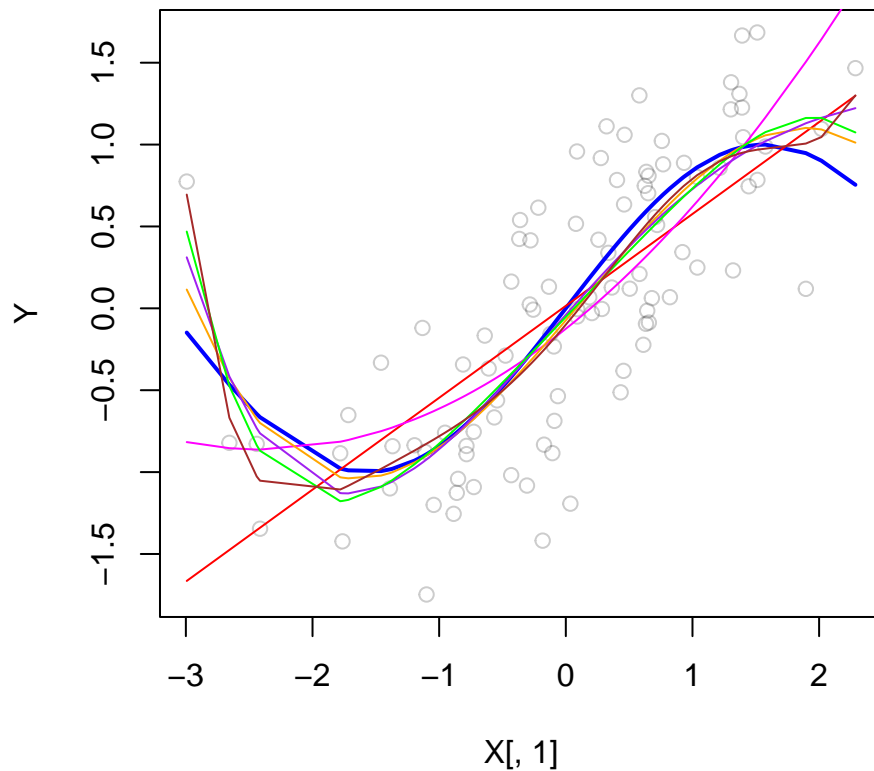
- the underdetermined problem:

```
try(ls_estimator <- solve(crossprod(X), crossprod(X,Y)))

## Error in solve.default(crossprod(X), crossprod(X, Y)) :
## System ist für den Rechner singulär: reziproke Konditionszahl = 5.84511e-18
```

- the bias-variance trade-off:

```
plot(X[,1], Y, col=rgb(0,0,0,0.2))
sX1 <- sort(X[,1])
points(sX1, sin(sX1), type="l", col="blue", lwd=2)
points(sX1, fitted(lm(Y ~ X[,1]))[order(X[,1])],
       type="l", col="red")
points(sX1, fitted(lm(Y ~ X[,1] + I(X[,1]^2)))[order(X[,1])],
       type="l", col="magenta")
points(sX1, fitted(lm(Y ~ X[,1] + I(X[,1]^2) + I(X[,1]^3)))[order(X[,1])],
       type="l", col="orange")
points(sX1, fitted(lm(Y ~ X[,1] + I(X[,1]^2) + I(X[,1]^3) +
                      I(X[,1]^4)))[order(X[,1])],
       type="l", col="purple")
points(sX1, fitted(lm(Y ~ X[,1] + I(X[,1]^2) + I(X[,1]^3) +
                      I(X[,1]^4) + I(X[,1]^5)))[order(X[,1])],
       type="l", col="green")
points(sX1, fitted(lm(Y ~ X[,1] + I(X[,1]^2) + I(X[,1]^3) +
                      I(X[,1]^4) + I(X[,1]^5) + I(X[,1]^6)))[order(X[,1])],
       type="l", col="brown")
```



- early stopping using a simple neural network:

```
library(dplyr)
library(keras)

neural_network <- keras_model_sequential()

neural_network %>%
  layer_dense(units = 50, activation = "relu") %>%
  layer_dense(units = 50, activation = "relu") %>%
  layer_dense(units = 1, activation = "relu") %>%
  compile(
    optimizer = "adam",
    loss      = "mse",
    metric    = "mse"
  )

history_minibatches <- fit(
  object      = neural_network,
  x           = X,
  y           = Y,
  batch_size  = 24,
  epochs      = 100,
  validation_split = 0.2,
  callbacks   = list(callback_early_stopping(patience = 50)),
  verbose     = FALSE, # set this to TRUE to get console output
  view_metrics = FALSE # set this to TRUE to get a dynamic graphic output in RStudio
)

plot(history_minibatches)
```

